A COMPOSED STATE-SPACE WORLD MODEL FOR ULTRASOUND-GUIDED SPINAL SURGERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Ultrasound-guided pedicle screw placement is studied as two disconnected problems: bring the probe to a standard imaging plane, and drill a planned corridor once you are there. The second is always evaluated with the probe *teleported* into place from ground truth, so no published number says whether the pair composes. We solve each problem independently with a state-space world model, then compose the two in sequence. Each is a learned readout of the tool's pose in a physically named frame, propagated by a small transition and planned by CEM-MPC over their respective objectives. The composition replaces the SonoGym simulator's privileged probe placement with a navigated one. In isolation each half is competitive: while learning its dynamics the drilling controller reaches 3.00 mm insertion and 2.80 mm lateral at 90% safety, beating both reported reinforcement-learning baselines on both position axes (their 11.9–12.3 and 5.07–5.42), and exact kinematics improve that again to $2.1/2.0$ mm; the probe controller reaches and holds the plane in 0.95 of episodes. Composed, safety falls from 90% to 0% and success to 0.00, and it fails in one place, not everywhere: drill orientation survives a 32° probe azimuth error while lateral offset does not. Instead of building a better navigator, we inject known amounts of probe error and measure how much the drilling controller absorbs. The pose estimator turns out not to degrade at the new probe position but to stop responding to its input, reporting the same large error wherever the probe is, which is what an estimator does when both its image and its own output are far outside the distribution it was fitted on. Post-training it at the position where it is actually used repairs this. The composed system then recovers most of what the teleport gave it: safety returns to 85% against the teleported arm's 90% with the exact propagator and to 75% at matched propagator.Auditing the forecasts the planner already computes and discards splits its 13.2 mm error into a propagator term and a perception term: two independent routes put the propagator's share at 9.2 and 9.4 mm, and correcting it leaves a perception-limited 3.8 mm. Two further probes put a hand-designed alignment cap, not the planner, in control inside the corridor. Three lessons about measuring such compositions, demonstrated here rather than established in general: one-channel tolerances do not compose, the joint failing where each channel alone passes; the tightest tolerance is on a probe axis neither task commands, failing by leaving the intervention undone while the safety metric still reads safe; and the controller's own introspective signals invert, the arm with the worst forecast controlling best and the one that fails completely proving the most self-consistent.

1 INTRODUCTION

A world model earns its name by answering counterfactuals, and the standard recipe answers them in a learned latent space: encode observations, fit a transition against the encoder's own next output, plan by rolling it forward (Watter et al., 2015; Ha & Schmidhuber, 2018; Hafner et al., 2019; 2025; Hansen et al., 2024). In ultrasound-guided intervention that recipe fails on both tasks, for two different reasons. The imaging probe is carried by a *second* arm, so the tool moves while the image barely changes. That is the obvious suspect, and it is not the one the measurement convicts: §3.1 bounds the action's authority over the propagated state for the probe task's transition design, measures it at exactly zero, and locates the collapse in the fitted transition rather than in the window it reads. The drilling task's transition sits outside that bound's hypothesis and its latent design fails for a statistical reason instead, so we propagate coordinates on *both* tasks but justify it differently on each. Propagating a named physical state instead is a well-populated line (Deisenroth & Rasmussen, 2011;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Chua et al., 2018; Nagabandi et al., 2020), and value equivalence makes precise why it suffices: the transition need only be right about what the objective reads (Grimm et al., 2020; Manuelli et al., 2020). The resulting system is competitive with the published baselines on both tasks while learning its dynamics.

The contribution of this paper is what happens when the two halves are put together. The surgery environment does not navigate its own probe: it teleports it to the standard plane from privileged ground truth and drives only the drill (Ao et al., 2025), so every drilling result in the literature, ours included, is conditioned on an oracle placement. Replacing that teleport with a navigated pose removes a privileged input, and that substitution, not the concatenation, is the experiment.

Contributions.

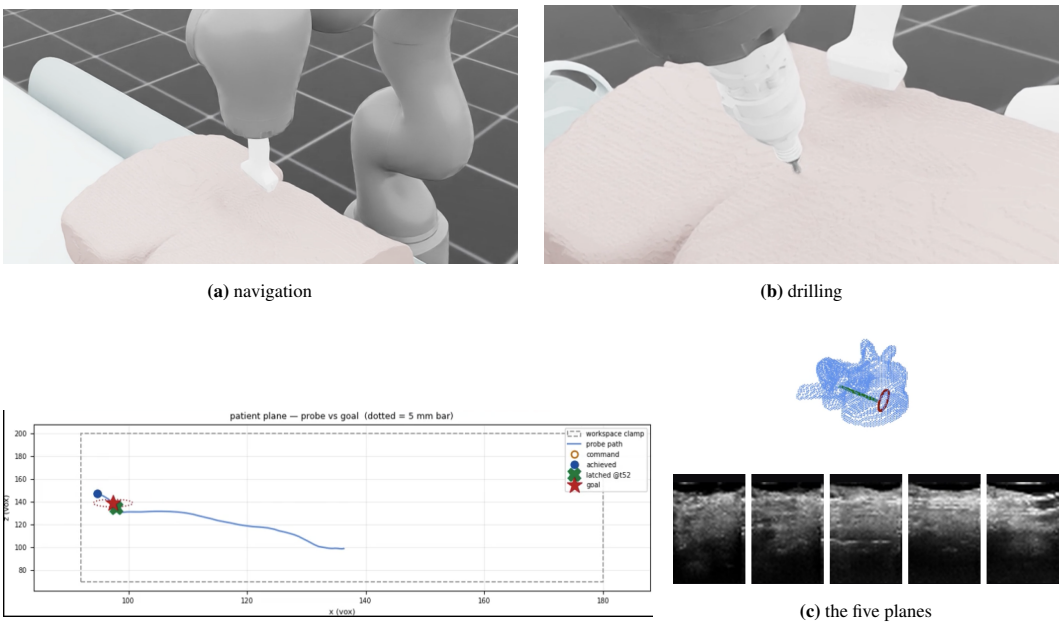
- A state-space world model, competitive with the baselines published with this simulator on both of its tasks, and on drilling better than both reinforcement-learning baselines on both position axes *while learning its dynamics* (§4). Exact kinematics improve on that again, so the learned propagator is close to free rather than free. One controller serves both tasks, and the designs differ only in how the state is rolled forward (§3, Appendix C).
- The first measurement of what happens when the two are joined and the probe is no longer placed from ground truth (§5), reported clause by clause so the failure can be located rather than only observed; and, because a composition run alone cannot say which half to fix, a controlled way to find out — inject known error into the downstream module's input and measure how much it absorbs (§6).
- The repair that measurement prescribes, carried out (§6): re-collecting the navigator against the drilling task's own probe placement rather than navigation's standard plane, and refitting its pose estimator there, restores the composed system to within one episode in twenty of the teleported one on safety.
- Two lessons about how tolerances should be measured, demonstrated on the joint error grid (§6): tolerances measured one channel at a time do not compose, failing where each channel alone is still acceptable (0.30 against an independence prediction of 0.56); and the tightest tolerance of all is on a probe axis *neither* task commands, which fails by leaving the intervention undone while the safety metric still reads safe.
- An audit of what the controller believes about itself (§7), built from forecasts its own planner computes at every step and discards: the forecast error separates into a perception term and a propagator term, the latter agreeing to 2% across two independent routes; the two signals that most look like competence checks both point the wrong way, the arm forecasting worst controlling most accurately and the arm that never succeeds proving the most reproducible across re-plans; and two further probes agree that inside the corridor a hand-designed alignment cap, not the planner, is in control.

2 Problem formulation

Both tasks are partially observed through synthesised ultrasound and share a simulator built on Isaac Lab (Ao et al., 2025; Mittal et al., 2023). **Navigation** asks a probe to reach *and hold* the L4 standard imaging plane from a single $1\times200\times150$ image with no proprioception; success requires $\leq 5\,\text{mm}$ and $\leq 10^\circ$ at the final step. On both tasks this is a demanding definition, and deliberately so: a conjunction evaluated at a single instant, so a run that drifts by the last step scores zero, as does one correct on three clauses and wrong on the fourth. Mean errors are therefore the gentler reading and SUCCESS the harsh one. **Drilling** asks a 6-DoF drill, with roll pinned so five are effective, to reach a target depth inside a pedicle corridor of half-length $\ell = 24.6\,\text{mm}$ and radius $r = 7.5\,\text{mm}$ (Youssef et al., 2024). The corridor geometry comes from the pre-operative plan and is given to the controller; the drill's pose *relative to it* is not, and must be inferred from five ultrasound slices and the probe-to-drill transform (Figure 1b). The simulator ships two renderers, a learning-based GAN (Isola et al., 2017) and a model-based ray-cast, which differ only in appearance; every result here uses the GAN renderer, and every condition is run for 20 episodes unless stated otherwise.

Writing d_t for depth along the corridor axis and ρ_t for lateral offset, the environment declares a state unsafe when the tip is committed but outside the corridor:

$$\textsc{unsafe}(s_t) \;=\; \underbrace{\big(d_t > -h\big)}_{\text{committed}} \;\wedge\; \neg\, \underbrace{\big(\rho_t < r + \epsilon \;\wedge\; d_t < \ell\big)}_{\text{inside the corridor}}, \qquad h = 50\,\text{mm}, \tag{1}$$



(a) navigation **(b)** drilling **(c)** the five planes

Figure 1: The two tasks running in the simulator, SonoGym on Isaac Lab. Top row, the scene camera; bottom row, the quantity each controller is judged on. **(a)** Navigation brings the probe to a standard plane; below, its path in the patient plane against the goal, inside the workspace clamp ($t = 156$, 12.9 mm out at 7.9°). **(b)** Drilling advances along the planned corridor; below, the anatomy with the corridor in red and the drill axis in green ($t = 180$, 4.2 mm insertion and 2.0 mm lateral against a corridor radius of 7.5 mm). **(c)** The five ultrasound planes the encoder receives, body surface at the top of each. The post-training of §6 retargets the navigator's goal from navigation's own standard plane to the probe pose the drilling task images from, and the composition then supplies (b)'s privileged probe placement out of (a).

a state-wise constrained decision problem (Altman, 1999), h being the simulator's own commitment depth. The predicate folds two clinically distinct events into one scalar, breaching the corridor wall and over-drilling past the target. They have opposite remedies, so we report them separately (Appendix I). It is also one-sided in depth: landing on target trips it, so insertion error and safe ratio are partly the same quantity read in opposite directions.

What the reported numbers are. Every metric is computed from the *true* pose, read before the step is applied, and is never shown to a controller. Writing T for the final step, drilling is scored on three error channels and one episode-level clause:

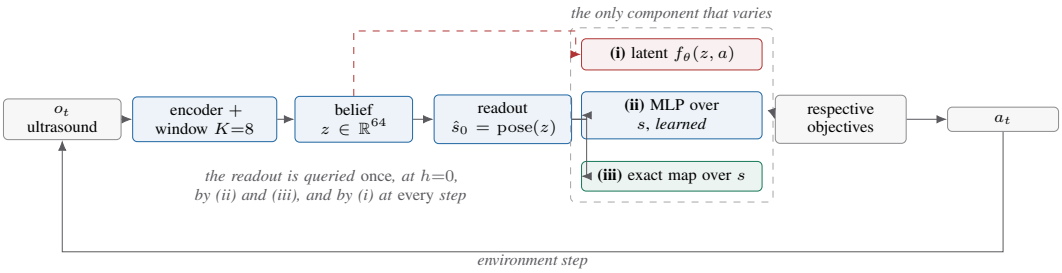
$$\underbrace{|d_T - \ell|}_{\text{INSERT}} < 5\,\text{mm}, \quad \underbrace{\rho_T}_{\text{SIDE}} < 5\,\text{mm}, \quad \underbrace{\arccos(\hat{u}_{\text{corr}} \cdot \hat{z}_{\text{tip}})}_{\text{ROT}} < 10^\circ, \quad \underbrace{\neg\,\text{UNSAFE}(s_t)\ \forall t}_{\text{SAFE}}, \tag{2}$$

that is, how far short of or past the target the tip stopped, how far off the corridor axis it sits, the angle between the corridor and the drill's own axis, and whether Eq. 1 was ever tripped. SUCCESS is the conjunction of all four, so a per-clause table always upper-bounds it and the lowest row names the axis a change would have to move. Navigation uses the same shape with FINAL POS and FINAL ROT.

3 CONTROLLER DESIGN

The controller is one stack, instantiated twice (Figure 2). A convolutional trunk (He et al., 2016) encodes the observation, together with the proprioceptive transform on drilling, to $e_t \in \mathbb{R}^{256}$; a small transformer over a window of $K{=}8$ embeddings produces a belief $z_t \in \mathbb{R}^{64}$; a readout takes z_t to the tool-to-goal relative pose, position plus a continuous 6-D rotation (Zhou et al., 2019). Perception is trained self-supervised, then supervised on privileged corridor labels with the encoder unfrozen, then refined by DAgger (Ross et al., 2011) on trajectories carrying held-direction perturbations, which manufacture the off-axis recovery states a well-behaved controller never visits. Perception is frozen whenever a propagator is trained, which is what makes an unperturbed arm a control.

The planner is CEM-MPC (Rubinstein, 1999; Williams et al., 2017) over a continuous corridor cost with a per-step Lagrangian safety penalty, re-planned every step (Appendix C). Safety is enforced twice and in different currencies: a graded penalty inside the cost, and an alignment gate implemented



	navigation	**drilling**
observation o_t	one US image, $1\times200\times150$	five US slices, $5\times50\times37$, + probe→drill transform $\in \mathbb{R}^7$
action a_t	$(\dot{x}, \dot{z}, \dot{\psi}) \in [-2, 2]^3$	6-DoF twist, roll $\equiv 0$, per-axis box
propagated state s	achieved $\oplus$ commanded $\oplus$ rot6d (12)	$p_{\text{DG}} \oplus$ rot6d (9)
what the cost reads	distance to the imaging plane	corridor depth, lateral offset, axis angle
written down in (iii)	clamped command integrator in the end-effector frame	rigid-body composition + the environment's action mask
safety layer	none	per-step $\lambda \sum_h \gamma^h \mathbb{1}[\text{UNSAFE}]$ + alignment gate
SUCCESS	≤ 5 mm *and* $\leq 10^\circ$ at the final step	≤ 5 mm insertion *and* lateral, $\leq 10^\circ$, never unsafe
not a sufficient state:	the command is clamped in absolute coordinates and rotated by a frame that follows the body surface, leaving an 8.37-voxel tracking lag	the action is a delta in the tip frame, so mapping it to a change in the corridor scalars requires the orientation

Figure 2: One controller, two instantiations, three propagators. Everything outside the dashed box is shared code; the comparison in §4 replaces one function and nothing else. The last row is why the propagated state is 12 and 9 numbers, not the 3 and 3 the objectives are written over: "interpretable" does not mean "minimal". The per-task pipelines, with module-level detail, are Figures 6 and 7.

as a *cap on the commanded forward action* (Wabersich & Zeilinger, 2021; Gros et al., 2020) rather than a cost term, because a cost-based gate reintroduces a discontinuity that a sampling optimiser refuses to cross. An earlier region-switched form parked the drill at the entry with 104 mm insertion error even given a perfect pose.

What each task commands. Navigation writes to the probe's own command vector, three numbers: two translations along the body surface and the probe azimuth ψ, a rotation about the world vertical, clamped to a workspace box. The probe's height and its two remaining rotations are not free; they follow the body surface at whatever position the command names. The observation is the ultrasound image, and the goal is never given: it is inferred as the achieved pose minus the readout's estimate of the offset to it. Drilling writes a 6-DoF delta in the tip frame with roll pinned, and reads five ultrasound slices plus the probe-to-drill transform. Its corridor geometry is given; its pose relative to that corridor is not.

What propagates the state across the horizon is a dynamic models over the physical pose and the action, fit by a discounted multi-step rollout loss along the recorded action sequence (Bengio et al., 2015) in millimetres and degrees, not normalised units. Its targets are *achieved* next poses, never the closed-form map's own output, since fitting a formula we already possess would be distillation. That distinction matters because the closed form is wrong in identifiable ways: the environment realises only $\approx 9\%$ of commanded tip translation per step and damps lateral and angular motion $5\times$ inside the corridor, which the analytic path hard-codes and a fitted one absorbs from data (Ajay et al., 2018; Zeng et al., 2020).

3.1 Why the state is coordinates and not the belief

All three propagators could carry the 64-dimensional belief; two carry the 9-dimensional pose instead, and that is a decision. Two reasons push the same way and they differ in kind.

On the probe task the softmax lets the action redistribute attention mass over a fixed window of observation embeddings but not add any, so its authority is capped by two factors: how far it can move the weights, and how much the embeddings it reweights differ. The second is ordinary. The first is what collapses, and it collapses in logit production: of the three ways a softmax gradient can

vanish, only that one carries any dynamic range, spanning $4{,}582\times$ against $1.08\times$ and $1.03\times$. Trained on that objective alone the sensitivity measures 0.000 against a plant response of 2.797 voxels per action unit (Appendix B), and the cause is the objective rather than the architecture or the optimiser: a term scoring the true action against negatives restores the factor by $6{,}376\times$ and costs a threefold loss of far-field readout accuracy, so the action-blind transition was the better minimiser of the predictive loss all along. Appendix D gives the bound and the measurement.

On the drilling task that bound does not apply, and the latent design fails anyway for a different, statistical reason: its transition behaves like a conditional-mean predictor, because a safety cap makes the executed action nearly a function of the state (§4). Attributing each arm's forward command to its own state channels makes that concrete: both coordinate arms are strongly monotone in depth, pushing less the deeper the tip already is (correlation between a channel's value and its attribution, -0.91 and -0.73), while the latent arm is flat at $+0.09$ — the same forward push whether the tip is 200 mm short or already past the target, a constant command wearing a state channel's clothes (Appendix H).

3.2 Training

Appendix A unfolds Figure 2 into the two tasks separately, as Figures 6 and 7. Both tasks are trained in two stages, with the details in Appendix B. Perception and the pose readout are fitted first on trajectories collected by a privileged controller, then the whole stack is refined on its own rollouts so that the readout sees the states the planner actually visits. The propagator, where it is learned, is fitted to *achieved* next poses read from the environment, never to the closed-form map's prediction of them.

4 Each half in isolation

Two labels on different axes. *Propagator* names what rolls the state forward and is the only component that varies among our designs: *latent* f_θ over the 64-dimensional belief, *MLP, learned* over the 9-dimensional corridor-frame pose, or *closed form*, the exact kinematics (Appendix C). *Oracle* names a privileged *input*, and there are two: an *oracle pose*, where the corridor-frame pose comes from ground truth rather than from ultrasound (Table 2, ceiling rows), and an *oracle probe*, where the environment teleports the probe to the standard plane rather than a navigator placing it (Table 3). They are independent, and every published drilling number is oracle-probe with an inferred pose. Only deployable rows are results; oracle rows are upper bounds that localise a gap.

Navigation. The probe controller reaches and holds the plane in 0.95 of episodes at 2.5 mm final position error (Table 1), inside the band reported for reinforcement learning on this task and well inside the imitation one. The two learned propagators split the axes between them: both cost most of the task on position, and both beat the closed form on *rotation*. That has an identifiable cause. The analytic rollout clamps commanded yaw and therefore believes an error floor of 8.02° that the achieved-pose metric does not, hence model fit on achieved motion never acquires that false belief.

Table 1: Probe navigation, against the baselines reported with the simulator (Ao et al., 2025). Errors are at the final step, the metric the simulator uses, with the best value the trajectory passes through in brackets, which separates failing to *reach* the plane from failing to *hold* it. The baseline rows are ranges, not point values: the paper reports bands, we quote the wider reading available to us, and SUCCESS is not reported there. Metric parity was verified against the released code; position matches exactly and rotation needed a wrap correction.

	method	propagator	FINAL pos ↓ (min)	FINAL rot ↓ (min)
	PPO (Schulman et al., 2017)	—	5–7	8–10
	imitation (Zhao et al., 2023)	—	< 16	< 12
LB	ours, latent-MPC (i)	latent f_θ	40.5 (31.0)	8.6 (6.6)
	ours, coord-MLP (ii)	MLP, learned	8.3 (0.8)	**5.2** (2.1)
	ours, analytic (iii)	**closed form**	**2.5** (**0.3**)	7.8 (4.6)

Drilling. The analytic arm reaches 2.1 mm insertion, 2.0 mm lateral, 2.5° and 90% safety, which beats both reported reinforcement-learning baselines (Schulman et al., 2017) on all three accuracy axes and is the only row that is simultaneously sub-3 mm on both position axes and $\geq 85\%$ safe. Imitation (Zhao et al., 2023) remains more accurate on insertion at lower safety. Both halves therefore clear the baselines published with the simulator, on the metrics those baselines are reported on. The ordering among our own propagators is not navigation's, though: here exact composition wins on every axis. Swapping the learned propagator for exact composition moves SUCCESS by two episodes, so the propagator's form is cheap, though not free. Rolling the learned *latent* transition never succeeds,

Table 2: Surgery drilling Every row of ours infers the pose from ultrasound except the oracle ceiling. SUCCESS is absent for the baselines because it is not reported with the simulator (Ao et al., 2025). *The latent row is the best of five latent designs* (§4): an MLA carrying the action's own key/value row alongside the window's, lowest of the five on insertion error. Its lower SAFE is not a safety regression but Eq. 1 reading an arm that actually approaches the target: the predicate is one-sided in depth, so an arm that stops short of the gate scores safe by construction.

	method	propagator	side ↓	insert ↓	rot ↓	safe ↑	SUCCESS ↑
LB	PPO (Schulman et al., 2017)	—	5.42	12.3	3.70	93.1	—
	PPO + safety filter	—	5.07	11.9	3.93	95.3	—
	ACT (Zhao et al., 2023)	—	5.34	1.62	1.01	86.0	—
	ours, latent-MPC (i)	latent f_θ	75.1	78.9	20.0	30.0	0.00
	ours, coord-MLP (ii)	**MLP, learned**	**2.80**	**3.00**	**4.20**	**90.0**	**0.75**
	ours, analytic (iii)	closed form	2.00	2.10	2.50	90.0	0.85
	oracle pose (ceiling)	*closed form*	*0.40*	*0.50*	*1.50*	*100.0*	*1.00*
	oracle pose (ceiling)	*MLP, learned*	*0.5*	*0.1*	*2.4*	*85.0*	*0.85*

though not for the reason of §3.1: this task's transition is outside that bound's hypothesis. Five architectures, four of them outside it, predict how much of a commanded advance occurs to within a factor of 1.7, which is what a conditional-mean predictor returns when the executed action is nearly a function of the state.

5 THE COMPOSITION

The two tasks place the probe at different offsets from the same vertebra, 45 mm and 90° apart, but both as vertebra-pose-plus-a-constant, so the retarget is exact and needs no refitting. It must be applied to the *readout* rather than to the state: shifting the state moves the goal and the rollout start together and cancels.

The substitution is one line in the environment's reset: the drilling controller consumes a probe pose, and we replace the teleported one with a navigated one, so only the error carried across that interface changes. Table 3 gives the arms, and Figure 3 decomposes what that costs clause by clause.

Table 3: The composition, before and after the fix. Rows 1–2 are the teleport this work removes: the probe is placed from ground truth. Rows 3–4 replace it with a navigated probe and change nothing else. Rows 5–6 post-train the navigator's readout against the drilling task's own probe pose (§6) and are the deployable result. Within each block the only difference is the propagator, so the last column isolates the propagator's contribution. *Handoff* is the error in the probe pose actually delivered, against ground truth. A closed-form navigator fails identically before the fix (19.5 mm insertion, SAFE = 0), so the failure is not the navigator's competence. All rows use a later on-policy round of the drilling checkpoint than Table 2: the two agree exactly on SAFE for both propagators and differ on SUCCESS, the brittle metric at this sample size.

probe / readout trained at	propagator	handoff (mm / °)	insert ↓	side ↓	rot ↓	safe ↑	SUCCESS ↑
oracle probe (reference)	*closed form*	—	*1.9*	*2.1*	*2.2°*	*90*	*0.85*
oracle probe (reference)	*MLP, learned*	—	*4.9*	*5.7*	*6.6°*	*90*	*0.20*
navigated, nav's own goal	closed form	10.33 / 32.32	19.5	18.6	5.4°	0	0.00
navigated, nav's own goal	MLP, learned	10.33 / 32.32	23.8	25.5	15.5°	5	0.00
navigated, drilling probe pose	**closed form**	**4.73 / 4.37**	2.4	3.4	2.6°	85	0.60
navigated, drilling probe pose	**MLP, learned**	**4.73 / 4.37**	**5.6**	**6.9**	**7.5°**	**75**	**0.20**

The navigation rollout reaches the right place and stops somewhere else. It passes through the correct pose in every episode, with a best approach of 2.09 mm and 0.46° and reach@5 mm of 1.00, then halts 10.33 mm and 32.32° away, because its own readout places the target 6.4 mm and ≈ 32° off. Scored against that estimate it succeeds in 0.600 of episodes; against ground truth, 0.000. The explanation is that the readout was trained with the probe near one angle and its command clamped, and at the drilling probe pose it is rotated 90° and shown a slice orientation nothing in its training data resembles. §6 shows this is a *saturation* and not a bias, and what repairs it.

The drilling rollout fails in one channel only. Safety goes 90% → 0% and success 0.85 → 0.00 from probe displacement alone, and the per-clause pass rates locate it exactly: rotation 0.75, insertion 0.30, lateral 0.05, safe 0.00. At the 32° azimuth error the navigator actually delivered, the drill's *orientation* clause still passes while its lateral offset does not, at 18.6 mm against a 7.5 mm corridor radius. Every episode reads off-axis and the drill parks at the entry gate, because the alignment gate

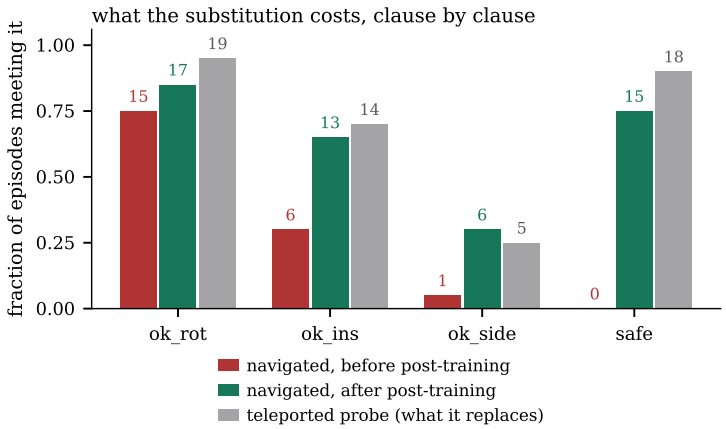


Figure 3: What the composition costs, clause by clause. Before post-training (red) the failure is not spread evenly: the orientation clause still passes at the delivered 32° probe azimuth error, while lateral offset falls to $1/20$ and safety to $0/20$. After (green), every clause returns to the teleported reference (grey). `ok_side` recovers least, and it is also the limiting clause for the teleported probe itself (6 episodes against 5). Numerals are episodes out of 20, and differences of one or two are not resolvable.

correctly refuses to advance a tip it believes is outside the corridor: the safety machinery behaves properly on a false premise. So a learned readout is accurate *locally to the goal it was trained at*, and composing two of them moves each outside its own accurate region.

6 What the handoff can absorb, and why the navigator could not deliver it

The composition says a 10.33 mm / 32.32° handoff destroys the task, but not which half is responsible, how much would have been acceptable, or what to repair. To answer the first two we make the displacement a controlled variable: a known offset δ is added to the probe pose before re-rendering, and the offsets are paired across magnitudes, so $\delta = 0$ reproduces the stock oracle arm exactly (Appendix F). For the third we instrument the navigator that produced the handoff.

Azimuth is a cliff, position is a slope (Figure 8). We read breakpoints off SAFE, the only clause with usable resolution at this sample size. Azimuth holds to 10° and breaks at 15°; position holds to 10 mm and breaks at 15 mm. Past 20° the lateral clause is exactly zero, because the corridor has left the imaging plane altogether and lateral offset becomes *unobservable* rather than merely inaccurate. Set against the measured handoff the azimuth is fatal and the position survivable, which tells a reader what to rebuild. One caveat: the $\delta = 0$ anchor scores only 0.15–0.20 with this checkpoint, so the drilling controller sets the headroom here and the handoff is measured against a low ceiling.

The channels do not compose. Sweeping one channel at a time defines a *rectangle* of admissible handoffs, position to 10 mm and azimuth to 10°, and the joint grid disagrees with that rectangle exactly where it matters (Figure 4b). At its corner the two marginals predict $\text{SAFE} = 0.56$ if the channels acted independently (Eq. 15); the measurement is 0.30 against a 0.75 anchor. No single cell carries that at this sample size, so the claim rests on the whole row: every cell at $\psi = 10^\circ$ falls below the independent prediction, and does so monotonically, at 0.87, 0.54 and 0.47 of it. Appendix F puts the single cell at about two sigma.

And the tightest axis is one neither task commands. The probe pose carries a roll the drilling environment draws for itself, which no action can set and which is neither the azimuth ψ nor the drill-to-corridor angle ROT of Eq. 2. Perturbing it is $4\times$ more damaging than perturbing ψ, and the two break the task in different places. Azimuth destroys *safety*. Roll leaves safety near 0.40 and destroys the *hole* instead, the drill parking short of the commitment gate rather than breaching it. Those two failures call for opposite remedies, and SAFE can only see the first of them (Appendix F).

The readout does not drift at the drilling probe pose. It saturates. We instrument the rollout with a *readout floor*, the smallest error the estimator reports anywhere on a trajectory. Before the fix that floor stands at 39.63 mm and 23.27°, against a true closest approach of 6.62 mm and 0.55°, and 26.4 voxels of it is approximately the 30-voxel retarget offset itself. The estimator is reporting "at navigation's own goal" wherever the probe actually is. The distinction matters for the repair, because a shifted estimate can be corrected whereas a constant carries no ordering to correct with.

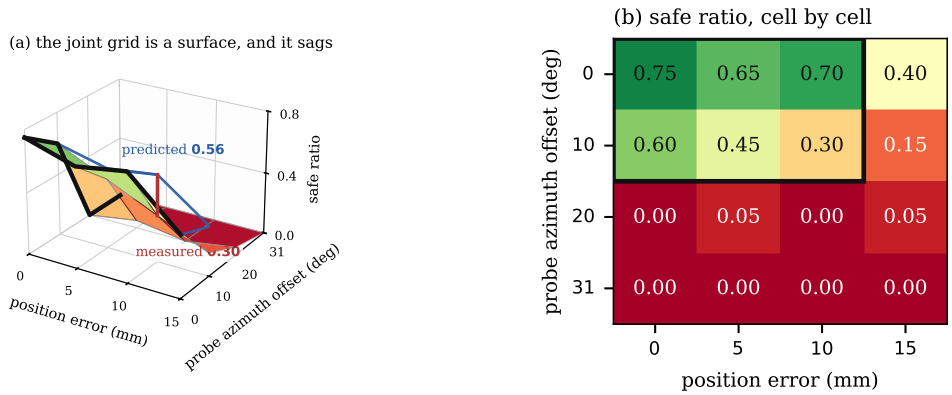


Figure 4: Both error channels together. (a) The 16 measured cells as a surface. The thick black edges are this grid's own single-channel margins and the blue mesh is what they predict jointly if the channels acted independently, drawn only where both are non-zero; the measured surface sits below that prediction across the near band. **(b)** The same cells at readable size, with the box the two breakpoints imply. Margins are read within this grid rather than across the separate sweeps of Figure 8, which agree on azimuth and differ by up to 0.30 on position. Learned pose-space propagator, GAN renderer, perturbation directions drawn once and reused at every magnitude so the grid is paired.

The raw errors agree, the angle error spanning only 32.03–32.69° across all twenty environments. No stopping rule repairs this: one built on the estimate never fires, and neither does a version given the privileged true distance, because position and angle are never simultaneously inside their bars. The close approaches of §5 are crossings of a sweep, not near-misses, so the fix has to be the estimator (Appendix E).

What we change is where the readout is trained. The estimator is out of distribution there in its own predicted output as well as in the image, and its response collapses. Imaging geometry compounds the problem. Across anatomies only the middle of the corridor falls inside the sampled volume, and the insertion target never does, so the corridor's position has to be inferred from surrounding tissue. We therefore leave the estimate alone at evaluation time and change what the navigator is collected against. This is stage two of the training protocol run again for a few DAgger rounds with one thing altered: its goal becomes the drilling task's own probe placement, and collection and the perception fit are re-run. The privileged controller that generates the data reads its goal from the environment, so no controller changes. The propagator is carried frozen, which is safe *by construction* rather than by measurement, since it consumes achieved pose, command and rotation only.

The composition then costs nothing relative to the teleport (Table 3). The handoff falls to 4.73 mm and 4.37° and the readout floor to 0.99 mm and 0.15°, reporting small when the probe is in fact 2.15 mm away; the deployable latch fires in every environment where it previously fired in none; and the navigated arm recovers the oracle-probe row on every clause. The handoff also stops predicting the outcome: success is flat across inherited-error bins (0.00 at 0–2 mm, 0.20 at 2–5, 0.25 at 5–10), correlation 0.204, weakly *positive* and therefore noise. *The privileged input is removed at no measurable cost.* One condition it forced: the retargeted goal lies 2.6 voxels *outside* the navigable range, so *both tasks' standard plane sits outside their own navigable workspace* and any navigated approach needs the range widened first.

7 What the controller believes about itself

Everything above measures the composition from outside. The same runs expose what the planner believes. Its elite set forecasts depth, lateral, angle and UNSAFE over the horizon for every candidate at every step and then discards them; and the forward command it commits to can be attributed exactly to the three state channels by enumerating all eight coalitions. Keeping the first and computing the second turns the controller's introspection into a measurement.

The forecast error decomposes, the propagator owns most of it. Only $h = 1$ is a clean propagator test: MPC executes one action then rebuilds. There, the depth forecast error is 0.1 mm from a true pose with the learned propagator, 4.0 mm once perception is learned, and 13.2 mm with the closed-form map, so perception costs 3.9 mm, the propagator 9.2. The closed-form map under-moves by a single scalar: 0.078 in depth and lateral, 0.071 in angle. Applying it in closed loop drops the error 13.2 → 3.8 mm, putting the propagator's share at 9.4. Two independent routes, 2% apart.

Neither signal that looks like competence is one (Figure 5). The arm with the worst forecast turns out to have the best control: the closed-form map reaches 41 mm at $h = 10$ and SUCCESS $= 0.85$, while the learned propagator reaches 5.5 mm and 0.75. This is objective mismatch (Lambert et al., 2020); what the audit adds is its mechanism here, a single scalar gain. Correcting the forecast does not close that gap either. It converts the closed-form arm into the learned one on every axis, with rotation moving $2.5 \rightarrow 4.4°$ onto the learned arm's $4.2°$. We rest the claim on those continuous channels, because the success difference is $17/20$ against $15/20$ and is not resolvable at this n. Re-planning the same belief $k = 16$ times measures self-consistency rather than correctness, and the arm that fails completely is the most reproducible of the three, so a runtime check that refused to advance when a plan would not reproduce would stay silent exactly where it is needed. Calibration cannot adjudicate either, starved as it is on the arms worth deploying, which are UNSAFE on 0.3% of windows (Appendix F).

Inside the corridor the planner is not the thing in control. The re-plan probe does establish one positive result. For both coordinate arms the alignment cap binds on 79–83% of re-plans and removes 84–87% of the forward command's spread, and a per-state attribution of the pre-cap command agrees that the command there is independent of the believed state (Appendix H). Two instruments, arrived at independently, put the drilling phase under a hand-designed gate.

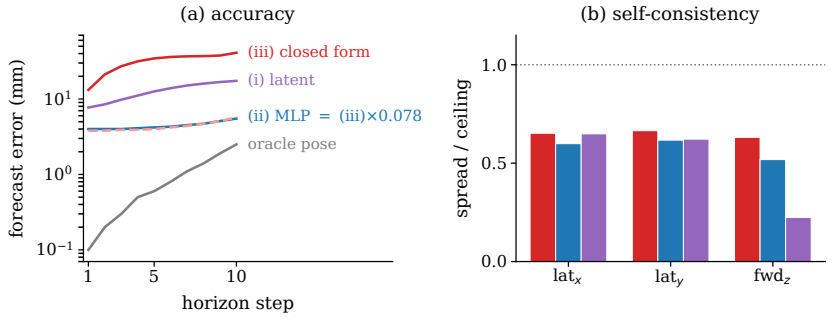


Figure 5: Two things the controller reports about itself. (a) Forecast error against horizon; the closed-form arm is the learned one divided by the gain. **(b)** Re-plan spread over the $\varepsilon = 0$ ceiling, ordered opposite to (a), which is the result. Closed-loop runs.

8 Discussion and Limitations

In SonoGym, the composition is the experiment the field's evaluation protocol hides. Reporting a drilling controller against a teleported probe is not wrong, but it conditions every number on an input that will not exist at deployment, and the conditioning is not small: the same controller goes from 0.85 to 0.00 when that input comes from a module instead of an oracle. What keeps this useful is that the failure is *localised* to one channel of one interface with a measured tolerance, so the fix could be specified instead of searched for. The honest headline is not "composition fails" but: *a privileged input can be removed at no measurable cost, provided the module downstream of it is trained where it will be used, and provided the interface is measured first.*

Limitations. The repair changes two things, not one: the post-trained arm had to be run with the navigation workspace widened to make its own goal reachable, which the arm before it did not need. We attribute the recovery to where the readout was trained, because what it removes is an estimator property and widening a clamp does not make an estimator report smaller numbers, but the control that would settle it has not been run. Corridor geometry is supplied, and the cost, gate and safety predicate are hand-designed in every arm, so "learned" refers to perception and the propagator, not the controller. Single patient, whose plant is a specification: exactly why the gap between a written-down map and achieved motion should widen on hardware, not close. Roll is a further instance: it never crosses the interface in simulation, because the handoff is the probe's three commanded numbers and the drilling environment draws its own roll at reset, so §6's roll result describes a channel not yet an interface variable here. On hardware, where probe roll is a mounting property fixed once by how the probe is clamped, it becomes one, and is what we would instrument first. Rendering is also a single operating point: both halves see the renderer they were trained on, so nothing here says whether the seam survives a change of image statistics, a second interface we have not opened. A readout that holds across patients is where this goes next.

ETHICS STATEMENT

This is a simulation study and supports no clinical claim. All results are produced in the SonoGym simulator (Ao et al., 2025), whose ultrasound is synthesised rather than acquired and whose anatomies are a fixed vendored set. The best deployable composed configuration reaches SUCCESS $= 0.60$ on one vertebral level of one anatomy family, which is far from any standard an interventional system would be held to. Nothing here is evidence that such a system is ready for use in or near a patient.
Two findings bear on how such systems should be evaluated, and both are cautionary. First, reporting a downstream controller against a privileged input overstates it substantially, and the overstatement is invisible in the reported metric, so that interface should be measured before composed performance is quoted. Second, we measure a failure a safety metric cannot see: under roll error the drill parks short of its target, remaining *safe* by the environment's own predicate while leaving the intervention undone, and a system reporting only a safety ratio would call that a success. We therefore report per-clause pass rates throughout, and recommend that evaluations of this kind decompose their aggregate and instrument the axes no component commands.

REPRODUCIBILITY STATEMENT

Code for every arm reported here — the controllers, the training pipeline, the perturbation harness and the evaluation scripts, with the configuration files that fix each run — together with the trained model weights and the dataset of recorded trajectories behind every table, will be released upon acceptance. The protocol is pinned in the appendices: the two task pipelines in Appendix A, the training protocol and its two stages in Appendix B, the planner objectives per arm in Appendix C, and the probe-error injection procedure behind the tolerance surface in Appendix F. The expressivity bound stated in §3.1 is proved in full in Appendix D, and the failure diagnosis of §6 is given in full in Appendix E. All results run in the SonoGym simulator (Ao et al., 2025) against its fixed vendored anatomy set.

AI USE STATEMENT

Among tasks requiring disclosure, we used large language models for one: feedback on the framing and interpretation of results. Among tasks with recommended disclosure, we used them to draft and edit the manuscript and to write plotting, data-processing, and harness code. No data here is LLM-generated; all trajectories and metrics come from the SonoGym simulator. The expressivity bound of §3.1 with its proof in Appendix D, the controllers and planner, and the hypotheses, experimental design, and evaluation protocol are the authors' own.
We have reviewed all AI-assisted work. Every AI-suggested framing was checked against the underlying measurements; the diagnosis of §6 rests on measurements the authors designed, ran, and verified. The authors read and ran all code and checked every figure against its source data. We take responsibility for the final content, including text, claims, and artifacts produced with generative AI.

REFERENCES

Anurag Ajay, Jiajun Wu, Nima Fazeli, Maria Bauza, Leslie P. Kaelbling, Joshua B. Tenenbaum, and Alberto Rodriguez. Augmenting physical simulators with stochastic neural networks: Case study of planar pushing and bouncing. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018.

Eitan Altman. *Constrained Markov Decision Processes*. Chapman and Hall/CRC, 1999.

Yunke Ao, Masoud Moghani, Mayank Mittal, Manish Prajapat, Luohong Wu, Frederic H. Giraud, Fabio Carrillo, Andreas Krause, and Philipp Fürnstahl. SonoGym: High performance simulation for challenging surgical tasks with robotic ultrasound. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, pp. 45013–45040, 2025.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, 2015.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Marc Peter Deisenroth and Carl Edward Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In *International Conference on Machine Learning (ICML)*, 2011.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: pure attention loses rank doubly exponentially with depth. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 2793–2803, 2021.

Christopher Grimm, André Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Sébastien Gros, Mario Zanon, and Alberto Bemporad. Safe reinforcement learning via projection on a safe set: How to achieve optimality? In *IFAC World Congress*, 2020.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning (ICML)*, 2019.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations (ICLR)*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, 2021.

Nathan Lambert, Brandon Amos, Omry Yadan, and Roberto Calandra. Objective mismatch in model-based reinforcement learning. In *Learning for Dynamics and Control (L4DC)*, volume 120 of *Proceedings of Machine Learning Research*, pp. 761–770, 2020.

Lucas Manuelli, Yunzhu Li, Peter Florence, and Russ Tedrake. Keypoints into the future: Self-supervised correspondence in model-based reinforcement learning. In *Conference on Robot Learning (CoRL)*, 2020.

Mayank Mittal, Calvin Yu, Qinxi Yu, Jingzhou Liu, Nikita Rudin, David Hoeller, Jia Lin Yuan, Ritvik Singh, Yunrong Guo, Hammad Mazhar, Ajay Mandlekar, Buck Babich, Gavriel State, Marco Hutter, and Animesh Garg. Orbit: A unified simulation framework for interactive robot learning environments. *IEEE Robotics and Automation Letters (RA-L)*, 8(6):3740–3747, 2023.

Anusha Nagabandi, Kurt Konolige, Sergey Levine, and Vikash Kumar. Deep dynamics models for learning dexterous manipulation. In *Conference on Robot Learning (CoRL)*, 2020.

Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. In *Advances in Neural Information Processing Systems 35*, 2022.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Reuven Y. Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodology and Computing in Applied Probability*, 1(2), 1999.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Kim P. Wabersich and Melanie N. Zeilinger. A predictive safety filter for learning-based control of constrained nonlinear dynamical systems. *Automatica*, 129, 2021.

Manuel Watter, Jost Tobias Springenberg, Joschka Boedecker, and Martin Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. In *Advances in Neural Information Processing Systems*, 2015.

Grady Williams, Andrew Aldrich, and Evangelos A. Theodorou. Model predictive path integral control: From theory to parallel computation. *Journal of Guidance, Control, and Dynamics*, 40(2): 344–357, 2017.

Salma Youssef, Jake M. McDonnell, Kielan V. Wilson, Luke Turley, Gráinne Cunniffe, Seamus Morris, Stacey Darwish, and Joseph S. Butler. Accuracy of augmented reality-assisted pedicle screw placement: a systematic review. *European Spine Journal*, 33(3):974–984, 2024. doi: 10.1007/s00586-023-08094-5.

Nikolay Yudin, Sergei Kudriashov, Alexander Gaponov, and Maxim Rakhuba. Pay attention to attention distribution: A new local lipschitz bound for transformers. *arXiv:2507.07814*, 2025.

Andy Zeng, Shuran Song, Johnny Lee, Alberto Rodriguez, and Thomas Funkhouser. Tossingbot: Learning to throw arbitrary objects with residual physics. *IEEE Transactions on Robotics*, 2020.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Josh Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, 2023.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems (RSS)*, 2023.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.